

디지털 콘텐츠의 URL 식별패턴 표준화에 관한 연구

A Study on the Standardization of URL Identifier Pattern for Digital Contents

김문정 중앙대학교 문헌정보학과
이두영 중앙대학교 문헌정보학과

Moon-Jung Kim and Too Young Lee
Dept. of Library and Information Science, Chung-ang University

아날로그 환경에서와 마찬가지로 디지털 환경에서도 디지털 콘텐츠 하나 하나에 고유 식별기호를 부여하여야 한다. 이러한 디지털 콘텐츠를 위한 식별기호로 IETF(Internet Engineering Task Force)에서는 URI(Uniform Resource Identifier)체계 하에 인터넷 자원에 대한 접근 메카니즘을 지정하는 URL (uniform resource locator)을 사용하고 있다. 그러나 도서관의 경우 각각 다른 OPAC(Online Public Access)시스템 환경 하에서 각각 다른 URL 식별 패턴을 사용하고 있기 때문에 동일한 자원을 검색하는데 있어서 문제가 되고 있는 것이 현실이다. 이러한 문제에 착안하여 본 연구는 디지털 콘텐츠에 대한 URL 식별구문패턴의 표준화 방안을 연구하고자 한다.

1. 서론

1.1 연구의 필요성 및 목적

정보기술의 혁신은 디지털 환경에서의 정보의 체계적 관리와 이용자로 하여금 디지털정보원에 대한 신속하고 정확한 접근을 필요로 하게 하였다. 최근 디지털정보원의 접근점에 대한 연구는 미국을 비롯하여 여러 선진국에서 활발히 이루어지고 있으며, 이러한 새로운 연구 결과들은 도서관이 자관의 정보원은 물론 도서관 외부에 산재되어 있는 정보원에 대한 접근까지도 가능하게 하고 있는 것이다.

전통적으로 도서관의 청구기호는 이용자의 서가 접근 과정에서 정보원의 객체 이동을 느끼지 못할 만큼 중요한 접근점으로 사용되고 있다. 이와 마찬

가지로 최근 전 세계적으로 이용이 확산되고 있는 인터넷 정보자원에 대한 식별 기호의 필요성이 크게 대두되고 있다. 그러나 현재까지는 디지털 정보 자원의 위치를 찾아내고 이에 접근하기 위한 식별 체제로 대부분 URL이 사용되고 있다. 이러한 URL 사용의 문제점으로는 도서관 이용자가 원격으로 각 도서관의 OPAC을 통해 정보원에 접근할 경우, 각각의 OPAC 시스템 내에서 각각 다른 정보자원에 대한 식별구문이 사용되고 있기 때문에 이용자는 동일한 콘텐츠에 대해 각기 다른 식별구문으로 검색해야 한다는 불편함은 물론, 검색의 신속성을 기하기 어렵다는 것을 들 수 있다.

이에 본 연구는 각각 다른 OPAC 시스템 안에서도 이용자가 신속하고 편리하게 디지털 콘텐츠를 검색할 수 있는 식별구문패턴의 표준화 방안을 제시하는데 그 목적이 있다.

1.2 연구의 내용 및 방법

본 연구는 디지털 콘텐츠에 대한 효과적인 식별 구분패턴의 표준화 방안을 도출하기 위해 다음과 같은 방법으로 연구를 수행하고자 한다.

첫째, 전통적인 아날로그 도서관 환경에서의 ISBN과 기존의 디지털 콘텐츠의 식별체제의 구조를 분석하고, 기존의 URL 식별구분에 따른 문제점을 분석한다.

둘째, 기존의 URL에 대한 분석은 도서관 OPAC 시스템에서의 웹을 통한 학술정보원을 대상으로 하였으며, 3개 사의 시스템을 대상으로 URL 식별구분을 분석하였다.

시스템 선정으로는 중앙대와 숭실대 도서관에서 사용하고 있는 (주)코아정보시스템의 AIMS, 부산대 도서관에서 사용하고 있는 (주)오름 정보시스템의 Vintage LAS, 서울대 도서관에서 독자 개발한 SOLARSnet으로 하였으며 이 3개의 OPAC환경에서의 URL 식별구분 패턴을 분석하고자 한다.

2. 이론적 배경

2.1 식별기호의 개념

식별기호는 사물의 식별을 위한 커뮤니케이션 도구로서 특히 콘텐츠의 축적과 검색에 핵심적 역할을 한다. 예를 들어 아날로그 환경에서 대표적인 ISBN 식별기호는 서적판매상과 출판사 사이에서 비즈니스를 위한 통신 수단으로 사용되고 있으며, 도서관에서는 자료를 식별하고 검색하기 위한 중요한 수단으로 사용되고 있다. ISBN과 같은 역할을 하는 ISSN 또한 출판사, 도서관, 연속간행물을 취급하는 도매상인들 사이에서 상업 행위를 촉진시키는 도구가 되고 있다. 또한 OCLC나 RLIN에서 부여한 식별기호의 서지적 활용도는 온라인 종합 목록 데이터베이스 구조에서의 중복적인 목록 데이터베이스의 발견과 통합을 용이하게 하는데 있다.

한편, 넓은 의미의 디지털 환경에서는 URI의

체계 하에 식별기호의 역할을 하는 URN과 메타데이터의 역할을 하는 URC, 실제 자원의 접근 메카니즘을 지정하는 URL로 구성되어 디지털 콘텐츠에 대한 정확한 접근을 가능케 하고 있다.

각각의 저작물에 대한 식별기호의 부여는 저작물의 저작권을 식별하는 강력한 행위 중의 하나으로써, 하나의 저작물에 대해 단 하나의 식별기호를 부여함으로써 저작물의 저작권을 보호하는 개념과 같다고 볼 수 있다.

2.2 아날로그 환경에서의 식별체제

식별기호는 각 정보원에 대한 명확한 구별을 제공하는 도구로서 ISBN은 대표적인 인쇄자료에 대한 식별번호라고 할 수 있다. ISBN은 출판사와 서적 판매상 간의 판매에 필요한 정보를 연계하는 핵심 역할을 해왔으며, 도서관에서 자료를 확인하는데도 중요하게 이용되어 왔다. 이외에도 출판사, 연속간행물 제공자(serial jobbers) 간의 업무를 활성화하는 역할을 수행하는 연속간행물 서명 식별기호로는 ISSN과 CODEN이 있다. ISSN은 연속간행물 체크인과 같은 도서관 내부적인 업무절차를 관리하는데도 중요한 요소로 사용되고 있다. 본 연구에서는 아날로그 환경에서의 식별기호로 대표적인 ISBN의 개념 및 구조만을 분석하고자 한다.

2.2.1 ISBN의 개념 및 구조

1) ISBN의 개념

ISBN은 국제적으로 표준화 된 규칙에 따라 출판물마다 각각의 고유한 번호를 부여함으로써 전 세계에서 발행되고 있는 모든 출판물의 식별을 명확하게 해주며, 발행된 출판물의 거래를 촉진시키고 서지적 정보유통의 정확성을 기할 수 있도록 고안되어 국제적으로 통용되고 있는 제도를 말한다. 즉, ISBN은 개개의 출판물마다 국제적으로 조정된 유일의 개별번호를 표시해서 국제적, 국내적인 유통이나 문헌의 정리, 검색 등에 이용하도록 한 시스템을 말한다.

2) ISBN의 구조

아날로그 환경에서의 대표적인 식별기호라 할 수 있는 ISBN의 구조는 10자리의 숫자로 구성되어 있으며 숫자가 시작되는 앞에는 반드시 ISBN이라는 문자를 표시하여야 한다.

10자리 숫자는 길이가 다른 4개의 부분으로 나뉘어 지는데, 각 부분 간의 표시는 하이픈을 넣어 구별하거나 띄어 쓰도록 되어 있다. 4개의 부분은 ① 그룹(국가 또는 지역) 기호, ② 출판자 기호, ③ 서명기호, ④ 대조 기호로 구성되어 있다.

① 그룹 식별번호

국제 ISBN 본부에서 배정하며, 이들 각 부분의 자리 수는 그 나라의 출판사 수 및 출판량에 따라 신축성 있게 사용할 수 있도록 가변적으로 운영되고 있다. 즉, 그룹 번호의 자리수가 적을수록 그 나라의 출판량이 많다는 사실을 의미하며 출판량이 적을수록 그룹 번호의 자리수가 늘어난다.

② 출판자 식별번호

이 부분은 그룹 내의 특정한 출판사를 표시한다. ISBN 시스템을 도입, 사용하고자 하는 출판자는 그룹 내에서 ISBN을 관리하는 국가코드관리위원회에 출판자 기호의 배정을 신청해야 한다.

③ 도서 식별번호

이것은 출판사번호의 길이에 따라서 그 출판사에 배당된 번호자리수의 범위 내에서 그 출판사가 간행한 특정 도서의 구체적 서명, 권, 판을 식별하도록 한 것이다.

④ 대조 기호

ISBN의 마지막 한 자리 숫자로, 컴퓨터에 입력할 경우 그 코드 번호가 바르게 부여되었는지를 자동으로 검색하기 위한 것으로 정해진 공식에 따라 계산하여 출판사에 표기한다.

2.3 디지털 환경에서의 식별체제

디지털 환경에서의 식별을 위해서도 아날로그 환경과 마찬가지로 디지털 자원 하나 하나에 고유 식별기호를 부여하여야 한다. 출판물의 경우에 식별기호는 상이한 자원을 구별해 주는 것을 핵

심 역할로 한다. 그러나 디지털 자원의 경우에 식별기호가 제대로 기능하려면 서로 상이한 자원을 구분함과 동시에 특정 자원의 물리적 위치를 지정해 주어야 한다. 현재 인터넷 자원의 식별기호로 쓰이는 URL은 일시적인 소장 위치에 관한 정보만 제공하므로 영속적인 접근성이 보장되어 있지 않고, 미리 사이트와 같이 상이한 URL로 표시되는 동일한 자원에 대해서는 아무런 통제가 이루어지지 않고 있다. 이와 같은 URL의 한계를 극복하기 위해서 IETF에서는 URI 체계를 개발해 오고 있다.

2.3.1 URI의 개념 및 구문규칙

1) URI의 개념

URI는 Uniform(통일), Resource(자원), Identifier(식별)이라는 세 가지 단어를 합친 개념이다. “통일”은 접근방법이 다른 다양한 유형의 정보식별기호를 동일한 방법으로 처리할 수 있으며, “자원”은 자료의 유형에 관계없이 식별이 가능한 “식별기호”를 갖는다는 의미를 나타낸다. 어떤 정보가 유일하게 식별됨으로써 비로소 시스템은 액세스, 갱신, 치환, 탐색 등이 가능하게 된다.

URI는 장소, 이름, 장소/이름으로 세분할 수 있다. URI의 부분집합인 URL은 정보를 정보의 이름이나 속성으로 식별하는 것이 아니라, “위치”(location)라는 기본적인 액세스방법으로 표현하여 식별한다. 또한 URI의 부분집합인 URN은 정보가 없어지거나 더 이상 이용할 수 없게 되더라도 유일하게 영구적으로 남을 수 있는 이름을 나타낸다.

2) URI의 구문규칙

URI의 구문은 scheme에 따라 다르지만 일반적으로 다음과 같은 형식을 갖는다

<scheme>://<authority><path>?<query>

이 구문은 확장성, 망라성, 인쇄가능성을 고려하여 정한 것이다. 각 요소는 예약문자를 이용하여

구분하게 되며, 필수부분이자 레이블인 <scheme>명을 제외하고 기타 부분은 없어도 무방하다. 나머지 부분은 <scheme>에 따르며, 형식에 따라서는 <authority>나 <query>가 없을 수도 있다.

<scheme>명은 소문자로 시작하며, 숫자 및 "+", ".", "-"를 포함할 수도 있다. <authority>는 naming authority로서 계층요소 중 최상위요소가 되며 일반적으로 인터넷서버가 되는 것이 보통이다. <authority>의 선두에는 "/"가 오며, "/", "?"가 오거나 URI가 끝나는 곳이 끝이 된다. <path>에는 scheme과 authority내에서 파일을 식별하기 위한 데이터가 포함된다. <path>내의 각 요소를 구별하기 위해 "/"를 사용할 수 있다. <path>내에서는 "/", ";", "=", "?"가 예약문자이다. <query>는 정보 자체를 나타내는 문자열이며, ";", "/", "?", ":", "@", "&", "=", "+", ",", "\$"가 예약문자이다.

URI가 잘 정의되어 있는 어떤 오브젝트 내에서 다른 오브젝트의 URI를 기술하는 경우, 두 URI의 일정 부분이 동일하다면 다른 오브젝트 URI의 그 부분을 생략형으로 사용할 수 있다. 이것을 부분형식(관계형식)(*partial form; relative form*)의 URL이라 한다. 즉, 어떤 위치의 파일을 읽고 있을 때, 액세스방법, 호스트, 포트번호, 디렉토리 등이 같은 다른 파일에 대해서는 부분형식 URL의 사용이 가능하게 된다.

2.3.2 URN의 개념 및 구문규칙

1) URN의 개념

URN은 인터넷 정보자원의 자원의 소재보다는 ISBN과 같이 그 실체를 규명하기 위해 할당된 간단한 텍스트 문자열이다.

1995년 12월 IETF 회의에서 발표된 제안서에 따르면, URN의 구성 요소는 1) 기명 전거(Name Authority) 시스템 식별 기호 2) 기명 전거 시스템 내에서 고유명의 역할을 하는 임의의 문자열의 두 가지로 정의된다.

2) URN의 구문규칙

기명 체계에 의해 정의되는 기명란(Name-space)은 URN을 지원하는데 필요한 구조와 서비스를 정의한다. 이로써 자원의 소장 위치가 바뀐 경우 변환 과정에서 이를 감지할 수 있게 되며, 기존의 포맷과 상이한 형태로 자원이 복제되었다라도 일관성 있는 서비스를 제공할 수 있다.

RFC 2141에 따르면 URN은 영구적이고 위치 독립적인 자원 식별기호로 사용되는 문자열로서 다음과 같은 구조를 가진다.

<URN> ::= "urn:" <NID> ":" <NSS>

여기서 <NID>는 namespace 식별기호이고, <NSS>는 namespace마다 고유한 문자열로서 실제 자원을 지칭한다.

2.3.3 URL의 정의

URL은 인터넷을 통해 이용 가능한 정보자원을 표현하는 함축된 문자열이다. 즉, URI의 하부 집합으로서 이미 네트워크 상에서 이용되고 있는 프로토콜을 이용하여 검색할 수 있는 객체, 즉 정보자원의 물리적 주소를 나타낸다. 이러한 정보자원의 위치를 기술하기 위한 스키마는 다양하게 존재하며 URL은 이들 스키마를 통해 정보자원에 접근하는 수단이다.

URL의 구문과 그 패턴에 대해서는 3장에서 더 자세히 다루고자 한다.

3. URL 식별패턴 분석

3.1 URL 식별구문 분석

URL은 글자, 숫자, 특수문자를 포함한 문자들로 구성된다. URL은 아스키코드 문자집합 가운데 인쇄 가능한 그래픽 문자만을 사용한다. 공백(space)이나 "<", ">", 따옴표(""), "#", "%", "{", "}", "|", "\", "^", "~", "[", "]", "" 등의 문자는

옥텟으로 바꾸는 인코딩이 필요 없이 URL내에서 그대로 코딩하면 된다. 일부 문자는 URL 스키마에서 특수 의미를 갖기 때문에 예약문자로 사용되는데 “:”, “/”, “?”, “;”, “@”, “=”, “&”와 같은 문자이다.

URL의 일반구문은 다음과 같다.

<scheme>:<scheme-specific-part>
<스키마명>:<스키마설명부>

스키마 이름은 일련의 문자로 구성되며 여기에는 소문자, 숫자, “+”, “.”, “-”을 사용할 수 있다. 대문자로 사용된 스키마 이름은 소문자와 똑같이 취급된다.

3.2 현재 OPAC 환경에서의 URL 구문

본 연구에서는 3개 도서관에서 각각 사용되고 있는 OPAC 시스템 즉, 중앙대와 송실대 도서관에서 사용하고 있는 (주)코아정보시스템의 AIMS, 부산대 도서관에서 사용하고 있는 (주)오름 정보

시스템의 Vintage LAS와 서울대 도서관에서 독자 개발한 SOLARSnet 이 3개의 OPAC환경에서의 URL 식별구문 패턴을 분석하고자 한다.

분석대상으로는 웹 데이터베이스에 수록되어 있는 학술저널을 선정하였으며, 선정된 저널은 “Accounting and Business Research”이며 저널의 ISSN은 0001-4788이다.

이 저널을 검색하였을 때 아래의 <표 1>과 같이 동일한 OPAC을 사용하고 있는 중앙대학교 도서관 시스템(AIMS)과 송실대학교 도서관 시스템(AIMS)에서는 웹 서버 주소만 다를 뿐 동일한 식별패턴으로 저널을 검색(링크) 할 수 있다. 하지만, 부산대학교 도서관(Vintage LAS)과 서울대학교 도서관(SOLARnet)의 경우는 각각 다른 OPAC 시스템을 사용하고 있기 때문에 중앙대와 송실대에서 검색되었던 저널 (ISSN:0001-4788)을 검색하려면 URL 패턴이 각 OPAC에 맞추어 다르게 기술되었음을 볼 수 있다.

<표 1> URL 패턴 조합에 의한 국내대학 OPAC 시스템 링크 예제

URL 패턴 조합에 의한 국내대학 OPAC 시스템 링크 예제	
1. 저널(Accounting and business research)에 대한 시스템별 링크	
중앙대 - AIMS	http://165.194.100.2/cgi-bin/mcu220?LIBRCODE=ATSL&USERID=*&SYSDB=R&EXPRS=IS=0001-4788
송실대 - AIMS	http://203.252.199.1/cgi-bin/mcu220?LIBRCODE=AM00&USERID=*&SYSDB=R&EXPRS=IS=0001-4788
부산대 - Vintage LAS	http://pulip.pusan.ac.kr/iplus/search/fulsearch/simp-relist.asp?fromyear=&toyear=&searchall=ON&maxsearch=100&ordercol=&ordermethod=asc&iStart=1&search_type=general&call_index_type=index&numtype=KO_ISSN&keyword=0001-4788
서울대 - SOLARSnet	http://solarsnet.snu.ac.kr/cgi-bin/srchproc.cgi?IPID=WEB&CMD_SW=SEARCH&SERVER=QUERY&REQPRG=http://solarsnet.snu.ac.kr/cgi-bin/series.cgi&SRCHIMG=&TITLE=종합검색%201945%20~&YEAR1=&YEAR=&WORD=&DATA=&SEARCH_KIND=SOLARSnetwork&UNIVERSITY=147.46.181.230&SEP_LIB_KIND=&COLLEGE_VAL=00&SEARCH=IS&TEXT=0001-4788

4. 현재 OPAC 환경에서의 문제점 및 개선방안

위의 URL식별 패턴을 볼 때 콘텐츠 식별 구문을 타 시스템에 적용하여 보았다.

예를 들어, 중앙대의 식별 구문인
"EXPRS=IS=issn_no"

와 같은 파라미터는 부산대의 OPAC이 해석을 못하는 것을 볼 수 있다. 이는 부산대의 OPAC에서는 같은 의미의 식별 파라미터를

"&num_type=KO_ISSN&keyword=issn_no"

와 같이 지정해야 하기 때문이다. 결국, URL 구문패턴이 표준화가 되지 않았기 때문이다.

그래서 위와 같이 콘텐츠 식별구문을 표준화 하였을 경우 <표 2>의 경우를 보면,
http://opac.web.server.addr/cgi-bin가 OPAC의 CGI-URL이고 ISSN=0014788이 콘텐츠 식별 파라미터가 된다. 그래서 이렇게 표준화가 되었을 경우에는 각 OPAC 시스템사의 표준화에 따라 System Specific Parameter 부분은 필요 없게 된다.

그러기 때문에 콘텐츠를 식별할 수 있는 URL 패턴의 표준화가 필요하다는 것을 알 수 있다.

<표 2> 콘텐츠 식별구문의 적용 예

<p>2. 콘텐츠 식별 구문을 타 시스템에 적용한 예 중앙대 형식을 부산대에 적용 http://lip.pusan.ac.kr/iplus/search/fulsearch/simp-relist.asp?fromyear=&toyear=&searchall=ON&maxsearch=100&ordercol=&ordermethod=asc&iStart=1&search_type=general&call_index_type=index&EXPRS=IS=0001-4788</p> <p>3. 콘텐츠 식별구문을 표준화 한 예 http://opac.web.server.addr/cgi-bin/stdsearch.cgi?issn=0001-4788</p>

5. 예상되는 결론

본 연구에서 나타난 것처럼 URL 식별구문을 표준화한다면, 이용자들이 서로 다른 OPAC을 대상으로 특정 콘텐츠를 검색하고자 할 때, 이용자들은 서로 다른 OPAC 환경에서도 동일한 URL 식별 구문을 통해 더 신속하게 원하는 정보원을 얻을 수 있을 것이다.

같은 형태의 URL식별구문 패턴을 만들기 위해서는 각 콘텐츠 (저널, 단행본, VOD 자료 등)에 따라 각각 식별 할 수 있는 코드와 식별구문 이름 등에 대한 프레임웍이 먼저 만들어져 각 OPAC 시스템사에서 제시되어야 할 것이다.

참고문헌

RFC 1738 Berners-Lee, T.. 1994 Universal Resource Identifier in WWW. [1999.6.30]
<http://info.internet.isi.edu:80/in-notes/rfc/files/rfc1630.txt>

Eric S. Curing up to Universal Resource Locators.[1997.7.7]
<http://www.mcli.dist.maricopa.edu/links/Curing-up-2-URLs.html>